

Variation sociolinguistique et réseau social : constitution et traitement d'un corpus de données orales massives

Aurélie Nardy, Hélène Bouchet, Isabelle Rousset, Loïc Liégeois, Laurence Buson, Céline Dugua et Jean-Pierre Chevrot



Édition électronique

URL : <https://journals.openedition.org/corpus/5561>

DOI : [10.4000/corpus.5561](https://doi.org/10.4000/corpus.5561)

ISSN : 1765-3126

Éditeur

Bases ; corpus et langage - UMR 6039

Référence électronique

Aurélie Nardy, Hélène Bouchet, Isabelle Rousset, Loïc Liégeois, Laurence Buson, Céline Dugua et Jean-Pierre Chevrot, « Variation sociolinguistique et réseau social : constitution et traitement d'un corpus de données orales massives », *Corpus* [En ligne], 22 | 2021, mis en ligne le 09 février 2021, consulté le 04 mars 2022. URL : <http://journals.openedition.org/corpus/5561> ; DOI : <https://doi.org/10.4000/corpus.5561>

Ce document a été généré automatiquement le 4 mars 2022.

© Tous droits réservés

Variation sociolinguistique et réseau social : constitution et traitement d'un corpus de données orales massives¹

Aurélie Nardy, Hélène Bouchet, Isabelle Rousset, Loïc Liégeois, Laurence Buson, Céline Dugua et Jean-Pierre Chevrot

1. Variation et réseau social en sociolinguistique

Dans le courant de la sociolinguistique variationniste, les langues sont conçues comme des systèmes hétérogènes et évolutifs du fait de leur dynamique interne, des contacts entre elles et de leur couplage avec l'organisation sociale, elle-même évolutive, composite et stratifiée (Labov, 1966 ; Weinreich, Labov & Herzog, 1968 ; Laks, 2013). Les usages langagiers des locuteurs sont contraints par des facteurs linguistiques et sociaux qui reflètent d'une part, les systèmes grammaticaux sous-jacents et d'autre part, l'organisation sociale des communautés dans lesquelles ils évoluent (Weinreich *et al.*, 1968 ; Coupland & Jaworski, 1997 ; Bayley, 2002). Un grand nombre de travaux, menés dans le sillage de ceux de Labov, a ainsi mis en évidence que l'usage des différentes variantes est contraint par des facteurs internes et externes à la langue. En ce qui concerne les facteurs extralinguistiques, les recherches ont mis en lumière des patrons de variation selon différents facteurs sociaux (profession, âge, genre, origine régionale) et contextuels (formalité de la situation d'interaction, thème de l'échange). Malgré la régularité des patrons de variation observés entre les différents groupes, les usages des locuteurs à l'intérieur de ces groupes ne sont pas tous homogènes. Deux locuteurs, de même âge, de même genre et de même milieu social, dans une situation d'interaction identique, peuvent produire des variantes différentes. C'est à partir de cette constatation qu'a été mise en place une approche plus microsociologique des phénomènes de variation s'appuyant sur le concept de réseau social.

[...] the network concept was developed for the very purpose for which we require it now – to explain individual behaviour of various kinds which cannot be accounted for in terms of corporate group membership (Milroy, 1987 : 135).

Une approche de la variation sociolinguistique à partir de l'étude du réseau social présente plusieurs avantages. Elle permet à la fois d'étudier des petits groupes de locuteurs qui ne sont pas opposables sur des critères macrosociologiques et d'appréhender la variation interindividuelle plutôt que la variation entre des groupes de locuteurs établis sur la base de catégories sociales prédéterminées (Milroy, 2002), tout en cernant la dynamique sous-jacente aux comportements langagiers variables des locuteurs. Par exemple, elle permet de rendre compte des patrons de variation observés chez les hommes et les femmes par des réseaux de socialisation différenciés. Les hommes, dont les usages sont plus vernaculaires, contractent généralement des liens plus denses et plus multiplexes dans le réseau local que les femmes qui manifestent des usages plus standards (Milroy, 1987 ; Labov, 2001 ; Chambers, 2009 ; Milroy & Llamas, 2013). Un dernier avantage, que nous reprenons à Labov (2001), est le fait que l'étude des usages langagiers des locuteurs au sein de leur réseau social permet de recueillir leurs productions dans leurs manifestations quotidiennes (conversations avec la famille, les amis, les collègues de travail). Cette approche plus microsociologique n'est pas incompatible avec une approche plus macrosociologique opposant différentes catégories de locuteurs. Il s'agit de deux niveaux d'analyse différents mais complémentaires puisque les réseaux sociaux des locuteurs ne sont pas indépendants de leur environnement social plus global qui lui-même affecte leur comportement langagier individuel. Notre position rejoint d'ailleurs celle de Milroy & Gordon (2003) qui avancent que l'examen des usages dans leurs conditions locales de production permet de comprendre les corrélations entre langage et catégories sociales plus globales.

Le réseau social d'un individu est défini comme étant l'ensemble des relations qu'il entretient avec les autres. Analyser ce réseau consiste à étudier les différentes structures et propriétés de ces relations (Milroy, 2002). Dans le cadre de l'approche par le réseau social en sociolinguistique, différents indicateurs ont été considérés ; les trois principaux étant des indicateurs sur les liens dans le réseau – densité et multiplicité (Milroy, 1987 ; Biichlé, 2018) –, des indicateurs d'intégration sociale (Milroy, 1987 ; Auer, Barden & Grosskopf, 2000 ; Kerswill & Williams, 2000 ; Beaulieu & Cichocki, 2002) et des indicateurs de statut sociométrique (Labov, 1972 ; Cheshire, 1982 ; Russell, 1982). Quels que soit la méthodologie adoptée et l'indice de réseau social choisi, les résultats des études menées chez l'adulte et l'adolescent aboutissent à une conclusion semblable : les liens forts renforcent l'usage des formes locales. À l'inverse, les liens faibles favoriseraient le changement linguistique.

Les enquêtes traditionnellement menées en sociolinguistique se sont appuyées sur les déclarations des enquêtés en saisissant les réseaux personnels des individus auprès d'échantillons ne dépassant pas la cinquantaine de participants. Cette approche égocentrée du réseau permet de rendre compte des influences sociales reçues par les individus mais elle ne permet pas de saisir l'influence réciproque des individus dans le réseau, ni de préserver l'écologie globale des liens sociaux au sein d'une communauté entière.

2. Le projet DyLNet

En vue de dépasser ces limites, dans le projet DyLNet² – *Dynamiques langagières, apprentissages linguistiques et sociabilité à l'école maternelle : apport des capteurs de proximité pour le recueil de données massives* – nous avons mis en place une démarche empirique fondée sur le recueil de données massives, profondes et longitudinales³ (Nardy, Fleury, Chevrot, Karsai, Buson, Bianco, Rousset, Dugua, Liégeois, Barbu, Crespelle, Busson, Léo, Bouchet & Dai, 2016). Il s'agit de cerner l'ensemble des relations dans une communauté – une école maternelle – au travers d'une approche en termes de réseau complet qui permet de conserver l'écologie des liens sociaux et de capturer la réciprocité des influences au travers du recueil dense, exhaustif et objectif des interactions sociales.

Le choix de l'école maternelle est un point de vue particulier mais il présente plusieurs avantages. Premièrement, ce terrain donne l'occasion d'étudier un phénomène rare : la formation d'un réseau social. Dans la vie de tous les jours, il est rare qu'un ensemble d'individus, qui ne se fréquentaient pas auparavant, se retrouvent régulièrement dans un même lieu dans lequel ils sont amenés à interagir et à tisser des liens. Deuxièmement, en France, la quasi-totalité des enfants de 3-5 ans est scolarisée. Du fait de cette généralisation de la scolarisation précoce et de la quantité de temps passée à l'école maternelle (24h par semaine, 864h par an), cette communauté est considérée comme une unité sociale signifiante dans les cultures occidentales (Hartup, 1984). Troisièmement, ce lieu permet d'observer la formation des réseaux affinitaires chez des individus dont la sociabilité est « neuve », les enfants entrant à l'école maternelle avec une expérience des pairs souvent limitée (à l'exception de ceux gardés en crèche ou faisant partie d'une fratrie). Contrairement au voisinage ou à la famille, la scolarisation met les enfants en position d'opérer des choix parmi l'abondance et la diversité des relations possibles et ce, souvent, pour la première fois de leur vie (Schaefer *et al.* 2010 ; Martin *et al.*, 2005). Quatrièmement, sur le plan développemental, la tranche d'âge 3-5 ans marque l'émergence des relations sociales avec les pairs du fait des modifications de leurs habiletés sociales et cognitives. C'est pendant cette période que les jeunes enfants passent d'une tendance à jouer seul ou à côté des autres à la mise en œuvre de jeux avec autrui accompagnés de comportements sociaux de recherche d'affiliation et d'engagement avec les pairs (Barbu 2003 ; Rubin *et al.*, 2006). Par rapport aux tout-petits, les enfants d'âge de maternelle affichent des réseaux plus grands, plus denses et plus organisés manifestant des amitiés réciproques avec des pairs particuliers plus stables (Johnson *et al.*, 1997 ; Strayer & Santos, 1996 ; Ladd, 1990). Finalement, sur un plan méthodologique, l'école ou la classe sont des collectifs dont les contours sont définis de manière exogène (Butts, 2008) par la frontière spatiale de l'institution et le temps scolaire (et non par les individus eux-mêmes) (Almquist, 2011). Ainsi, pour toutes ces raisons, la période de scolarisation en maternelle apparaît comme une occasion unique d'étudier la dynamique du développement interactionnel social et langagier précoce (Martin *et al.*, 2005).

Puisque les individus ne sont pas uniquement des réceptacles des processus de diffusion/circulation linguistique dans le réseau mais sont aussi des agents de ces processus qui possèdent des caractéristiques qui leur sont propres, nous disposons pour chacun d'eux d'informations sociocognitives, linguistiques et sociodémographiques. Enfin, les données longitudinales nous permettent de suivre

l'évolution des usages langagiers en fonction de la structure du réseau social et d'appréhender d'éventuelles interactions bidirectionnelles entre réseau et langage.

Notre approche consiste à suivre pendant toute la période de scolarisation en maternelle – 3 ans – près de 200 individus (enfants et intervenants pédagogiques) d'une école maternelle socialement mixte. Tous sont équipés une semaine par mois de capteurs sans fil qui enregistrent, toutes les 5 secondes, les proximités entre individus. De plus, les usages langagiers des enfants et des adultes sont enregistrés grâce à des micros intégrés aux capteurs. À plusieurs reprises durant les trois années, les enfants passent une série de tests psycho- et sociolinguistiques. Enfin, leur profil social est établi grâce à un questionnaire rempli par leur famille.

Cette collecte aboutira à la constitution d'une base de données caractérisée par un volume important, de la variété (parole, contacts sociaux, informations sociodémographiques, tests) et un flux d'entrée rapide résultant de la capacité des capteurs à enregistrer en continu, caractéristiques fondamentales de ce que sont les *Big Data* (De Mauro, Greco & Grimaldi, 2016).

La mise en place de ce recueil de données massives et variées a nécessité de relever plusieurs types de défis : prise en compte des aspects éthiques et du traitement des données personnelles⁴, développement de matériel *ad hoc*, traitement et analyse des données, puis stockage et conservation de celles-ci. Parmi ces défis, la question du recueil et du traitement des données audio massives est centrale pour le sociolinguiste. Dans les paragraphes suivants, nous présentons la méthodologie mise en place.

3. Recueil et traitement des données orales massives

Dans un premier temps, nous exposons les choix techniques que nous avons opérés afin de concevoir, en collaboration avec une entreprise spécialisée dans la collecte et le traitement de données mobiles, le dispositif d'enregistrement audio embarqué. Dans un second temps, nous présentons les différents pré- et post-traitements des signaux audio permis par le dispositif d'enregistrement pour enfin exposer la procédure de transcription en cinq étapes que nous avons mise en place.

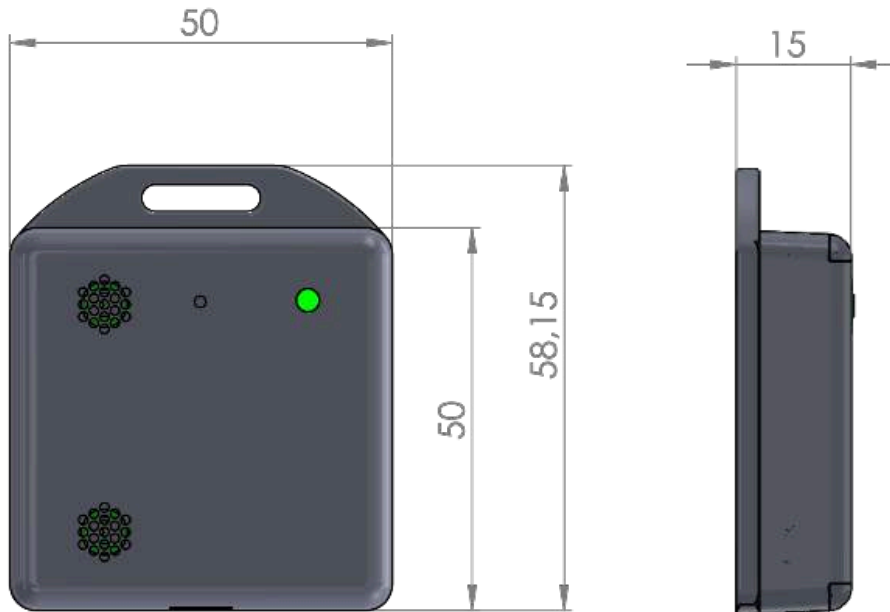
3.1. Recueil de données orales massives : dispositif d'enregistrement embarqué

Afin de recueillir les usages langagiers des participants durant tous leurs déplacements dans l'enceinte de l'école, nous avons opté pour un dispositif d'enregistrement embarqué que nous avons conçu afin qu'il satisfasse à la fois aux spécificités de notre terrain d'enquête et à nos besoins pour la recherche. La majorité des participants sont de jeunes enfants (âgés de 2 ans et demi à 6 ans et demi) qui sont enregistrés en continu une semaine par mois lors de leurs différentes activités quotidiennes à l'école (classe, cour de récréation, sport). Le dispositif d'enregistrement devait donc ne pas gêner ceux qui le portent et répondre à des exigences d'innocuité et de facilité d'équipement. Dans le même temps, il devait inclure une capacité de stockage suffisante (24 heures d'enregistrement chaque semaine), satisfaire des critères de qualité des signaux enregistrés et enfin permettre leur horodatage.

En partenariat avec une entreprise de la région Auvergne-Rhône-Alpes, nous avons développé des boîtiers de 58,15 x 50 x 15 mm (hauteur x largeur x profondeur) portés

au col grâce à une pince bretelle qui, en plus d'enregistrer toutes les 5 secondes les proximités entre individus, incluent deux microphones sur la face avant (cf. Figures 1 et 2).

Figure 1. Dispositif d'enregistrement embarqué



Les signaux recueillis sont enregistrés au format *.wav* (encodage 16-bit signed PCM, stéréo, 32kHz) sur une carte SD de 32 gigas. Par ailleurs, le système inclut un dispositif d'horodatage.

3.2. Traitement de données orales massives

3.2.1. Dispositifs de pré- et post-traitements automatiques des signaux

Le nombre de participants ainsi que la masse de données récoltées ($\approx 30\,000$ heures/an⁵) impliquent de relever deux principaux défis liés au traitement des fichiers audio recueillis. Le premier nécessite de reconnaître la voix du porteur du micro dans un environnement multi-locuteurs et bruyé. Dans le cas des voix des jeunes enfants, la tâche est encore plus complexe puisque, contrairement aux adultes, les fréquences fondamentales moyennes des deux sexes ne sont pas différentes (Weinberg & Bennett, 1971 ; Busby & Plant, 1995 ; Lee, Potamianos & Narayanan, 1999). Le second défi implique d'optimiser la tâche de transcription, particulièrement fastidieuse, afin de la rendre la moins chronophage possible tout en récoltant des données de qualité permettant de répondre aux objectifs de la recherche. À cette fin, différents dispositifs de pré- et post-traitements automatiques des signaux audio ont été appliqués aux enregistrements recueillis.

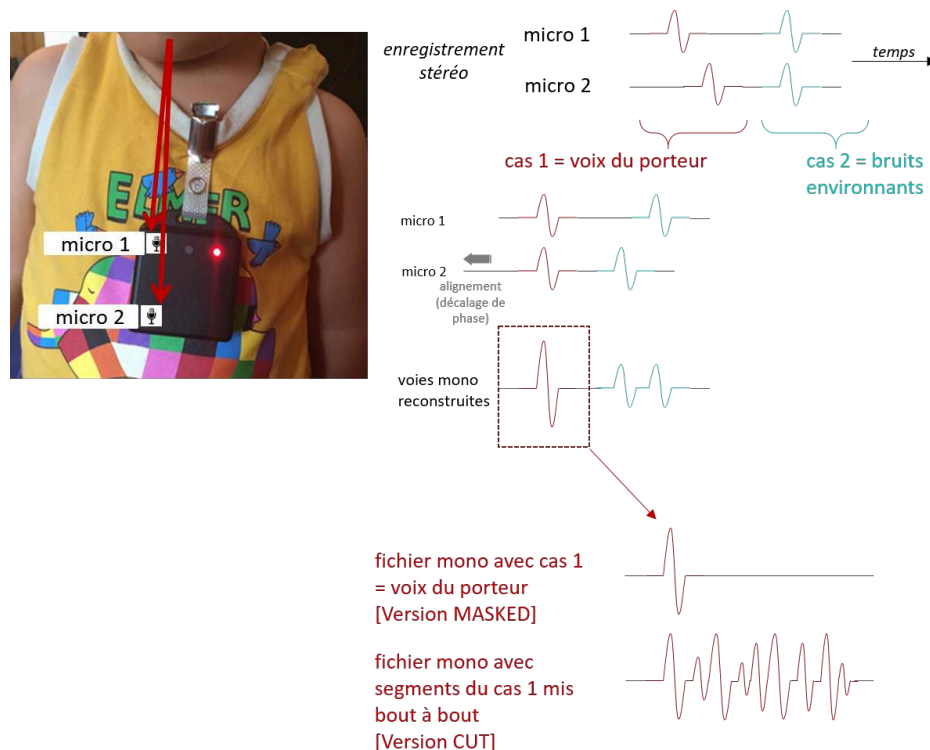
Une fois les enregistrements audio extraits des capteurs (pour chaque participant, deux fichiers d'environ 3h par jour correspondant à l'enregistrement en continu de la matinée et de l'après-midi), un premier traitement est appliqué afin de procéder à leur découpage en tranches horaires. Ce procédé permet de conserver l'horodatage initial et

génère des fichiers d'une durée d'une heure maximum. Par exemple, pour une matinée, nous obtenons 4 fichiers : 8h34-9h, 9h-10h, 10h-11h, 11h-11h18. Pour chacun de ces fichiers, nous disposons de différentes versions : originales et post-traitées.

Concernant les versions stéréo originales, elles se présentent d'une part au format *.wav* (version à partir de laquelle se font les transcriptions) et d'autre part au format *.flac* (format de compression sans perte de qualité pour le stockage et la conservation⁶).

À partir de la version stéréo originale, un post-traitement automatique des signaux est appliqué par un filtrage spatial (*beamforming*) réalisé sur la base du décalage physique entre les deux micros intégrés à chaque capteur. La figure 2, ci-après, illustre le principe général de ce post-traitement automatique des signaux.

Figure 2. Illustration du post-traitement automatique des signaux par filtrage spatial



Comme l'illustre la photo sur la Figure 2, lorsque le porteur du boîtier parle, sa voix arrive d'abord au micro 1 (plus proche de sa bouche) puis quelques millièmes de secondes plus tard au micro 2. Les autres sons environnants (autres que ceux de la voix du porteur) arrivent quant à eux plus ou moins en simultané aux 2 micros ou avec un décalage différent de celui suscité par la voix du porteur. Dans le cas de la voix du porteur, le post-traitement aligne les signaux identiques reçus par les 2 micros en décalé (ce décalage est maîtrisé puisque nous connaissons la distance exacte entre les deux micros). Ensuite, les signaux reçus par le micro 1 et le micro 2 sont sommés pour créer une voie mono qui permet de générer 2 versions post-traitées des fichiers originaux :

- une version *MASKED* dans laquelle seuls les segments audio identifiés par le post-traitement comme émanant du porteur du capteur sont présents, le reste des segments correspondant aux bruits environnants étant remplacés par du silence. Ainsi, la version *MASKED* conserve le déroulement temporel de l'enregistrement ;

- une version *CUT* dans laquelle les segments identifiés comme parole du porteur sont mis bout à bout.

Ce post-traitement automatique, qui dépend en grande partie de la position du locuteur par rapport aux micros du capteur, n'est pas parfait : certaines prises de parole du porteur peuvent ne pas être détectées ou certains bruits, comme le frottement du boîtier par exemple, peuvent être détectés. Il est toutefois précieux pour deux raisons principales. Premièrement, il est utile lors de la sélection des fichiers audio à transcrire puisque le ratio entre la durée du fichier stéréo original et celle du fichier *CUT* correspondant permet d'avoir une idée approximative du temps de parole du porteur sur la durée totale du fichier. Deuxièmement, il est utile lors de la phase de transcription pour cibler la voix du porteur. Sur la version *MASKED*, un script⁷ qui segmente le signal en silence/parole est lancé depuis *PRAAT* (Boersma & Weenink, 2013). À l'issue de ce processus, nous récoltons un fichier *.TextGrid* qui contient les temps de silence et de parole. Ce fichier est ensuite ouvert dans le logiciel de transcription alignée *ELAN* (Wittenburg, Brugman, Russel, Klassmann & Sloetjes, 2006) avec le fichier audio stéréo original, affichant ainsi les segments temporels qui indiquent au transcripteur les passages lors desquels le porteur prend la parole.

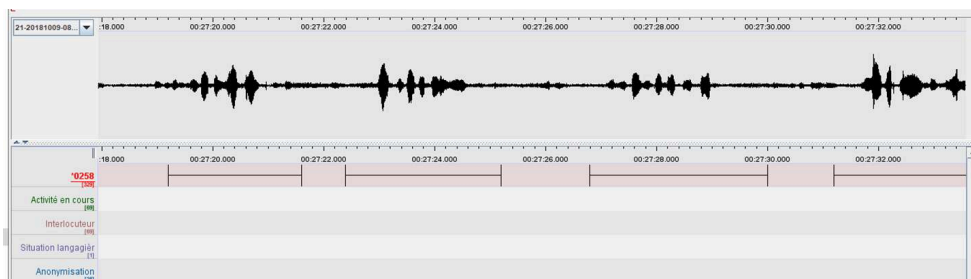
3.2.2. Procédure de transcription et d'annotation en 5 étapes

Conscients de la complexité de l'activité de transcription/annotation et de l'attention qu'elle requiert (Baude & Dugua, 2011 ; Hriba, Baude & Dugua, 2011), nous avons mis en place une procédure en cinq étapes successives visant un double objectif : optimiser le rendement du transcripteur tout en lui fournissant les outils nécessaires à l'obtention de données de qualité. Elle est mise en œuvre à partir du logiciel *ELAN* qui permet de réaliser une transcription alignée sur le signal et d'annoter différents types d'informations. Les différentes étapes permettent au transcripteur de faire plusieurs écoutes de l'enregistrement lors desquelles son attention est attirée sur des éléments spécifiques.

Étape 1 : Ajustement des segments temporels placés automatiquement et annotation d'éléments contextuels (1^{re} écoute)

Quand il ouvre le fichier audio stéréo original et le fichier *.Textgrid* dans *ELAN*, le transcripteur visualise les bornes placées automatiquement ainsi que la grille d'annotation qui lui a été fournie (cf. Figure 3).

Figure 3. Vue d'ELAN avec les bornes placées automatiquement et la grille d'annotation



Lors de cette 1^{re} étape, sa tâche est d'écouter les segments repérés par le post-traitement automatique présenté plus haut, de les ajuster au signal et de procéder à l'annotation de différents éléments contextuels nécessaires à l'étude de la variation tels

que l'activité en cours (contexte d'enregistrement : en classe, en récréation, en sport, situation indéterminée), l'interlocuteur (un ou plusieurs enfant(s), l'enseignant ou l'ATSEM de sa classe, un autre adulte, la classe entière, interlocuteur indéterminé ou parole auto-adressée), et la situation langagière (codage qui ne concerne que certaines prises de parole spécifiques comme les récitations ou la lecture à voix haute).

Cette 1^{re} étape permet au transcripateur de focaliser son attention sur un nombre d'éléments limité et, dans le même temps, de se familiariser avec l'enregistrement avant l'étape suivante de transcription à proprement parler.

Étape 2 : Transcription des paroles (2^e écoute)

Lors de ce second passage, le transcripateur écoute les segments bornés à l'étape précédente et les transcrit en suivant les conventions établies pour le projet DyLNet.

Étape 3 : Vérification et complétion de certains codages (3^e écoute ciblée)

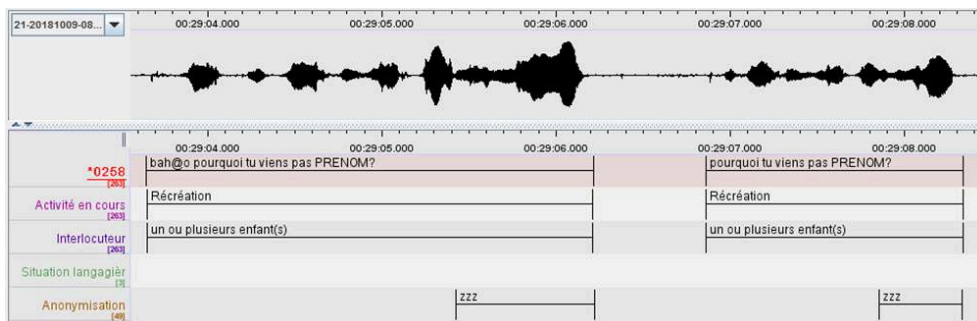
La troisième étape consiste en une réécoute ciblée de certains passages transcrits afin que le transcripateur vérifie et/ou complète ses codages de plusieurs variables sociolinguistiques du français⁸. Pour ce faire, il dispose de plusieurs recherches prédéfinies qui, lancées depuis ELAN, permettent de pointer les segments dans lesquels elles apparaissent. Les variables sont les suivantes :

- production ou non du 'u' de *tu* devant voyelle ou *h* muet
- production ou non du 'l' dans 'il(s)' et production complète ou non de 'il(s)'
- production ou non du schwa dans les 9 monosyllabiques les plus fréquents : *je, ce, de, le, me, ne, que, se, te*
- production ou non de /ʁ/ et /l/ en position post-consonantique finale

Étape 4 : Pré-anonymisation des signaux

En conformité avec les engagements pris envers les participants et les instances éthiques, tous les éléments permettant d'identifier des participants ou des lieux ne sont pas transcrits et sont remplacés par des suites de caractères en majuscules (PRENOM, LIEU). Lors de cette 4^e étape, le transcripateur lance une requête qui recherche toutes les suites de caractères en majuscules afin de lui permettre de créer le segment correspondant au signal audio dans la ligne « Anonymisation » de la grille d'annotation. Cette procédure permettra ultérieurement de lancer un script qui ira brouter les segments identifiés dans cette ligne pour anonymiser les enregistrements. Une illustration est donnée sur la Figure 4 ci-dessous.

Figure 4. Vue d'ELAN avec la grille d'annotation complétée



Étape 5 : Relecture orthographique et formelle de la transcription

La dernière étape de la procédure consiste à exporter la transcription vers un éditeur de texte afin de procéder à une relecture orthographique et formelle et de corriger les éventuelles coquilles et erreurs orthographiques dans le fichier de transcription.

Afin de fiabiliser et d'homogénéiser les transcriptions en vue d'un traitement automatique, un script a été développé (Fagot & Hamrouni, 2020). À partir d'un corpus de 508 000 mots transcrits dans le cadre de notre étude, nous avons extrait la liste des mots dont le nombre d'occurrences était supérieur à 3. L'examen de cette liste nous a conduits à définir une liste de remplacements automatiques visant l'uniformisation des transcriptions comme, par exemple, remplacer les occurrences de *t-shirt* par *tee-shirt*, supprimer les espaces avant ou après les apostrophes, etc.

4. Apports de l'étude à l'analyse variationniste

Au moment où nous rédigeons cet article, le recueil de données est toujours en cours. Dans cette dernière partie, nous présenterons les potentialités offertes par ce type de recueil pour affiner, automatiser et systématiser l'analyse variationniste des relations entre réseau social et variations sociolinguistiques.

En ce qui concerne les aspects méthodologiques, les travaux antérieurs se fondaient essentiellement sur les déclarations des enquêtés ou des indicateurs globaux de réseaux évalués par le chercheur. L'utilisation de la technologie des capteurs de proximité permet d'établir le(s) réseau(x) social(aux) des individus à partir de la fréquence objective de leurs interactions sociales avec une forte densité temporelle (une mesure toutes les 5 secondes, 1 semaine par mois pendant 3 ans). De même, le recueil en continu des productions langagières permet d'appréhender l'éventail de leurs usages sociolinguistiques dans des situations d'interaction et face à des interlocuteurs divers. En outre, du fait de l'automatisation du recueil de données et des capacités de traitement de données massives des sciences des réseaux, la taille des groupes peut être plus importante que dans les recherches antérieures. À notre connaissance, peu de recueils de données dans le domaine de la sociolinguistique de l'oral réunissent ces qualités de granularité, d'exhaustivité et de dynamité.

Grâce à la collaboration multidisciplinaire entre des linguistes et des chercheurs en science des réseaux, les relations entre réseau social et variations sociolinguistiques seront examinées dans une perspective dynamique. Il s'agira en effet de décrire la coévolution entre dynamique des réseaux sociaux (les changements dans les liens sociaux au sein de la communauté) et dynamique du langage dans les réseaux (les influences entre locuteurs et la modification de leurs habiletés langagières) en détectant les corrélations temporelles entre le(s) réseau(x) d'interactions et les usages sociolinguistiques au travers de deux approches : la première, traditionnelle dans les études menées en sociolinguistique, testant l'influence sociale sur les usages langagiers et la seconde, plus originale, examinant l'homophilie langagière sur le(s) réseau(x) de relations sociales (Aral, Muchnik & Sundararajan, 2009). De plus, nous aborderons ces questions à différents niveaux de description de la structure sociale : le niveau collectif (l'école dans son intégralité, la classe ou encore les groupes sociaux d'origine), le niveau intermédiaire (les groupes affinitaires) et le niveau individuel (les individus et leurs caractéristiques particulières).

En ce qui concerne l'influence sociale sur les usages langagiers, nous testerons si le changement de distance sociale entre les individus prédit la distance linguistique entre

eux. Si des individus restent dans la même communauté entre deux temps d'observation, nous caractériserons la probabilité que la distance linguistique entre eux diminue entre ces deux temps d'observation ou dans ceux qui suivent. De la même façon, nous nous demanderons si la distance sociale entre des individus au temps T ou son évolution entre T et T+n prédit leur distance linguistique à T+p (avec $p \geq n$).

L'approche consistant à mesurer l'influence du langage sur la structure sociale testera si la distance linguistique entre des individus prédit leur distance sociale. Si des individus appartiennent au même groupe linguistique (établi à partir de leurs usages sociolinguistiques) au temps T, nous déterminerons si la probabilité qu'ils soient dans la même communauté augmente à T+n. Similairement, nous verrons si la distance linguistique entre des individus au temps T ou son évolution entre T et T+n prédit leur distance sociale à T+p.

Notre étude longitudinale, d'une durée de 3 ans, permettra aussi l'examen de la propagation – ou, à l'inverse, de l'absorption – d'un changement ponctuel à la fois dans la structure sociale et dans le langage en mesurant, quantifiant et modélisant les processus d'influence à long terme. En effet, un élément que notre étude a l'opportunité de saisir de façon très fine est la perturbation des classes due aux rentrées scolaires, notamment lors des deuxième et troisième années de recueil de données, car elles remanient partiellement la composition des classes et donc potentiellement les communautés qui s'étaient mises en place durant l'année précédente. Nous pourrions ainsi évaluer l'effet de l'intégration dans une nouvelle communauté. Une communauté a-t-elle systématiquement un rôle homogénéisateur, d'absorption des changements linguistiques ou, au contraire, peut-elle accueillir les usages des nouveaux membres et les amplifier en les diffusant ?

BIBLIOGRAPHIE

Almquist Y. (2011). « The school class as a social network and contextual effects on childhood and adult health : findings from the Aberdeen Children of the 1950s cohort study », *Social Networks* 33(4) : 281-291.

Aral S., Muchnik L. & Sundararajan A. (2009). « Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks », *Proceedings of the National Academy of Sciences* 106(51) : 21544-21549.

Auer P., Barden B. & Grosskopf B. (2000). « Long-term linguistic accommodation and its sociolinguistic interpretation : evidence from the inner-German migration after the Wende », in Mattheier K. (dir.), *Dialect and Migration in a Changing Europe*. New York, Peter Lang, 79-98.

Barbu S. (2003). « Stability and flexibility in preschoolers' social networks : a dynamic analysis of socially directed behavior allocation », *Journal of Comparative Psychology* 117(4) : 429-439.

Baude O. & Dugua C. (2011). « (Re)faire le corpus d'Orléans quarante ans après : quoi de neuf, linguiste ? », *Corpus* 10 : 99-118.

- Bayley R. (2002). « The quantitative paradigm », in Chambers J. K., Trudgill P. & Schilling-Estes N. (dir.), *The handbook of language variation and change*. Malden, Oxford : Blackwell, 117-141.
- Beaulieu L. & Cichocki W. (2002). « Le concept de réseau social dans une communauté acadienne rurale », *Canadian Journal of Linguistics* 47(3-4) : 123-150.
- Biichlé L. (2018). « Intégration, réseaux sociaux et représentations langagières de migrants en France », *Langage et société* 163(1) : 33-56.
- Boersma P. & Weenink D. (2013). *Praat : doing phonetics by computer*. <http://www.fon.hum.uva.nl/praat/>.
- Busby P. A. & Plant G. L. (1995). « Formant frequency values of vowels produced by preadolescent boys and girls », *Journal of the Acoustical Society of America* 97(4) : 2603-2607.
- Butts C. T. (2008). « Social network analysis : a methodological introduction », *Asian Journal of Social Psychology* 11(1) : 13-41.
- Chambers J. K. (2009). *Sociolinguistic theory : linguistic variation and its social significance*. Oxford, Cambridge : Wiley-Blackwell.
- Cheshire J. (1982). *Variation in an English dialect: a sociolinguistic study*. Cambridge : Cambridge University Press.
- Coupland N. & Jaworski A. (1997). *Sociolinguistics : a reader and coursebook*. Basingstoke : Macmillan Press.
- De Mauro A., Greco M. & Grimaldi M. (2016). « A formal definition of Big Data based on its essential features », *Library Review* 65(3) : 122-135.
- Fagot J. & Hamrouni Y. (2020). *Outil d'analyse automatique de transcriptions issues d'un corpus oral*. Projet Professionnel du Master 2 Industries de la langue, Univ. Grenoble Alpes.
- Hartup W. W. (1984). « The peer context in middle childhood », in Collins W. A. (dir.), *Development during middle childhood*. Washington DC, National Academy Press, 240-282.
- Hriba L., Baude O. & Dugua C. (2011). Transcrire : la norme, la variation et le linguiste. *Colloque du CerLiCO : Transcrire, écrire, formaliser 2*. Orléans.
- Johnson J. C., Ironsmith M., Whitcher A. L., Poteat G. M., Snow C. W. & Mumford S. (1997). « The development of social networks in preschool children », *Early Education and Development* 8(4) : 389-405.
- Kamper H. (2017). *Unsupervised neural and Bayesian models for zero-resource speech processing*. PhD thesis. <https://arxiv.org/pdf/1701.00851.pdf>
- Kerswill P. & Williams A. (2000). « Creating a new town koine : children and language change in Milton Keynes », *Language in Society* 29(1) : 65-115.
- Labov W. (1966). *The social stratification of English in New York City*. Washington DC : Center for Applied Linguistics.
- Labov W. (1972). *Language in the inner city : studies in the Black English Vernacular*. Oxford : Basil Blackwell.
- Labov W. (2001). *Principles of linguistic change, vol. 2 : social factors*. Oxford : Blackwell.
- Ladd G. W. (1990). « Having friends, keeping friends, making friends, and being liked by peers in the classroom : predictors of children's early school adjustment ? », *Child Development* 61(4) : 1081-1100.

- Laks B. (2013). « Why is there variation rather than nothing ? », *Language Sciences* 39 : 31-53.
- Lee S., Potamianos A. & Narayanan S. (1999). « Acoustics of children's speech : developmental changes of temporal and spectral parameters », *Journal of the Acoustical Society of America* 105 : 1455-1468.
- Martin C. L., Fabes R. A., Hanish L. D. & Hollenstein T. (2005). « Social dynamics in the preschool », *Developmental Review* 25(3-4) : 299-327.
- Milroy L. (1987). *Language and social networks*. Oxford : Blackwell.
- Milroy L. (2002). « Social networks », in Chambers J. K., Trudgill P. & Schilling-Estes N. (dir.), *The handbook of language variation and change*. Oxford : Blackwell, 549-572.
- Milroy L. & Gordon M. (2003). *Sociolinguistics : method and interpretation*. Oxford : Blackwell.
- Milroy L. & Llamas C. (2013). « Social networks », in Chambers J. K. & Schilling N. (dir.), *The Handbook of Language Variation and Change (second edition)*. Malden : Blackwell Publishing, 407-427.
- Nardy A., Fleury É., Chevrot J.-P., Karsai M., Buson L., Bianco M., Rousset I., Dugua C., Liégeois L., Barbu S., Crespelle C., Busson A., Léo Y., Bouchet H. & Dai S. (2016). « DyLNet – Language Dynamics, Linguistic Learning, and Sociability at Preschool : Benefits of Wireless Proximity Sensors in Collecting Big Data (<https://dylnet.univ-grenoble-alpes.fr/>) », <ANR-16-CE28-0013>: <https://hal.archives-ouvertes.fr/hal-01396652>.
- Rubin K. H., Bukowski W. & Parker J. G. (2006). « Peer interactions, relationships, and groups », in Eisenberg N. (dir.), *Handbook of child psychology (6th edition) : social, emotional, and personality development*. New York : Wiley, 571-645.
- Russell J. (1982). « Networks and sociolinguistic variation in an African urban setting », in Romaine S. (dir.), *Sociolinguistic variation in speech communities*. London : Edward Arnold, 125-140.
- Schaefer D. R., Light J. M., Fabes R. A., Hanish L. D. & Martin C. L. (2010). « Fundamental principles of network formation among preschool children », *Social Networks* 32(1) : 61-71.
- Strayer F. F. & Santos A. J. (1996). « Affiliative structures in preschool peer groups », *Social Development* 5(2) : 117-130.
- Weinberg B. & Bennett S. (1971). « Speaker sex recognition of 5- and 6-year-old children's voices », *Journal of the Acoustical Society of America* 50(4) : 1210-1213.
- Weinreich U., Labov W. & Herzog M. I. (1968). « Empirical foundations for a theory of language change », in Lehmann W. P. & Malkiel Y. (dir.), *Directions for Historical Linguistics : a symposium*. Austin : University of Texas Press, 96-195.
- Wittenburg P., Brugman H., Russel A., Klassmann A. & Sloetjes H. (2006). « ELAN : a professional framework for multimodality research », in, *Proceedings of the Fifth International conference on Language Resources and Evaluation*. 1556-1559.

NOTES

1. Cette recherche est financée par l'Agence Nationale de la Recherche <ANR-16-CE28-0013>.
2. Site web : <https://dylnet.univ-grenoble-alpes.fr/>.
3. Ce projet s'inscrit dans une collaboration multidisciplinaire réunissant des sociolinguistes (LIDILEM, Univ. Grenoble Alpes) et des chercheurs en science des réseaux (équipe DANTE, Laboratoire de l'Informatique du Parallélisme, INRIA Grenoble Rhône-Alpes, ENS de Lyon).

4. Validation du protocole de l'étude par le *Comité Opérationnel d'Évaluation des Risques Légaux et Éthiques* (COERLE) de l'INRIA (avis favorable n° 2017-014), déclaration auprès de la *Commission Nationale de l'Informatique et des Libertés* (CNIL) pour la mise en œuvre du projet de recherche (avis favorable n° 2017_0980683), déclaration de la base de données auprès de la CNIL pour la constitution d'une base de données de recherche (avis favorable n° 2019_0980683).

5. Dans le cadre de ce projet, nous ne visons pas la transcription intégrale de l'ensemble des enregistrements audio recueillis. En collaboration avec des chercheurs en traitement du signal, nous envisageons une exploitation automatique des signaux notamment au travers du *key word spotting* (Kamper, 2017). Il s'agira ainsi d'entraîner un système à la reconnaissance automatique de séquences sonores correspondant à des mots et séquences de mots spécifiques puis, en cas de taux de fiabilité satisfaisant, de généraliser la recherche des mots et séquences cibles à l'ensemble du corpus.

6. Une évaluation faite sur une partie de nos fichiers audio indique un gain de 65% en termes d'espace de stockage par rapport aux mêmes fichiers en version *.wav*. Un autre avantage du format *.flac* est de permettre aisément la conversion vers le format *.wav*.

7. Nous remercions S. Rossato du Laboratoire d'Informatique de Grenoble (Univ. Grenoble Alpes) pour la mise à disposition de ce script.

8. Ces variables pour lesquelles cette procédure de vérification a été mise en place sont celles qui sont annotées directement lors de l'étape de transcription. D'autres variables sont également traitées mais elles font l'objet d'autres procédures de codage, ultérieures à l'étape de transcription. Concernant la variable de la négation en français, nous avons développé un script permettant, dans les fichiers transcrits, de détecter et annoter automatiquement les cas où la première particule est présente et ceux où elle est absente (évaluation du script en cours). Pour la liaison facultative, la procédure se fait en deux étapes. Dans un premier temps, un script fondé sur des règles lexicales, repère des contextes de suite de deux mots répondant à ces règles et insère un codage dans les fichiers transcrits. Dans un second temps, deux codeurs experts réécoutent tous les segments dans lesquels apparaît ce codage. Ils valident ou non le fait que ce soit un contexte de liaison facultative et, le cas échéant, indiquent si celle-ci est réalisée ou non.

L'ensemble de ces variables seront analysées dans le cadre du projet ainsi que d'autres permettant d'évaluer le niveau langagier enfantin comme la longueur moyenne des énoncés, la diversité lexicale, etc.

RÉSUMÉS

Nous présentons une étude originale en cours visant la compréhension des relations entre variations sociolinguistiques et réseau social. Sa démarche empirique repose sur le recueil de données sociales et langagières massives et longitudinales au sein d'une école maternelle. Environ 200 individus (enfants et adultes) sont équipés une semaine par mois pendant 3 ans de capteurs qui enregistrent en continu à la fois leurs interactions verbales et leurs contacts sociaux. Dans cet article, à visée principalement méthodologique, nous exposons les dispositifs mis en place pour le recueil et les procédures mises en œuvre pour la transcription et l'annotation de données orales massives. Enfin, nous présentons les potentialités offertes par ce type de recueil pour affiner, automatiser et systématiser l'analyse variationniste des relations entre réseau social et variations sociolinguistiques.

We present an outstanding ongoing study designed to understand the relationships between sociolinguistic variation and social networks. Its empirical approach is based on the collection of massive and longitudinal social and linguistic data within a preschool. Approximately 200 individuals (children and adults) are equipped one week per month during 3 years with sensors that continuously record both their verbal interactions and their social contacts. In this article, primarily methodological in nature, we present the devices set up for data collection and the procedures implemented for the transcription and annotation of massive oral data. Finally, we discuss the opportunities provided by this type of dataset to improve, automatize and systematize the variationist analysis of the relationships between social network and sociolinguistic variation.

INDEX

Mots-clés : variation sociolinguistique, réseau social, données massives, transcription

Keywords : sociolinguistic variation, social network, big data, transcription

AUTEURS

AURÉLIE NARDY

Univ. Grenoble Alpes, LIDILEM

HÉLÈNE BOUCHET

Univ. Grenoble Alpes, LIDILEM

ISABELLE ROUSSET

Univ. Grenoble Alpes, LIDILEM

LOÏC LIÉGEOIS

Univ. Paris Diderot, Sorbonne Paris Cité, CLILLAC-ARP et LLF

LAURENCE BUSON

Univ. Grenoble Alpes, LIDILEM

CÉLINE DUGUA

Univ. Orléans, LLL UMR7270

JEAN-PIERRE CHEVROT

Univ. Grenoble Alpes, LIDILEM